DATA ANALYSIS I

Types of Attributes
Sparse, Incomplete, Inaccurate Data

2019-20

Sources

- Bramer, M. (2013). *Principles of data mining*. Springer. [12-21]

- Witten, I. H., Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques [3rd Ed.]*. Morgan Kaufmann.[39-60]

- Zaki, M. J., Meira Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press. [1-3]

Concept

- Four basically different styles of learning:
  - *Classification learning.* The learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.
  - *Association learning.* Any association among features is sought, not just ones that predict a particular *class* value.
  - *Clustering.* Groups of examples that belong together are sought.
  - *Numeric prediction.* The outcome to be predicted is not a discrete class but a numeric quantity.
- The thing to be learned is a *concept*. The output produced by a learning scheme is a *concept description.*

Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff

Rules

- Part of a structural description of the weather data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

Problems

- In the weather problem there are $4 \times 4 \times 3 \times 3 \times 2 = 288$ possibilities for each rule.
- Not all rules are meaningful…
- We can restrict the rule set to contain no more than 14 rules (14 examples in the training set).
- How many combinations do we have?

Quality of Rules

- To find all rules, we have to execute the rule-induction procedure once for every possible combination of attributes, with every possible combination of values, on the right side. That results in an enormous number of rules.

- We can use rules with high

  - *support* (based on the number of instances in the rule)

  - *confidence* (based on correctly predicted instances in the rule)

Support and Confidence

$$supp = \frac{r}{N}$$

where *r* is the number of instances in the rule and *N* is the number of instances in the dataset.

$$conf = \frac{c_{max}}{r}$$

where *r* is the number of instances in the rule and c_{max} is a maximum number of correctly classified instances in the rule.

Weather Data (Numeric)

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Weather Data (Numeric Class)

Outlook	Temperature	Humidity	Windy	Play Time
Sunny	85	85	false	5
Sunny	80	90	true	0
Overcast	83	86	false	55
Rainy	70	96	false	40
Rainy	68	80	false	65
Rainy	65	70	true	45
Overcast	64	65	true	60
Sunny	72	95	false	0
Sunny	69	70	false	70
Rainy	75	80	false	45
Sunny	75	70	true	50
Overcast	72	90	true	55
Overcast	81	75	false	75
Rainy	71	91	true	10

Contact Lens Data

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Iris Data

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
…					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
…					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
…					

Labor Negotiations Data

Attribute	Type	1	2	3	…	40
duration	(number of years)	1	2	3		2
wage increase 1st year	percentage	2%	4%	4.3%		4.5
wage increase 2nd year	percentage	?	5%	4.4%		4.0
wage increase 3rd year	percentage	?	?	?		?
cost-of-living adjustment	{none, tcf, tc}	none	tcf	?		none
working hours per week	(number of hours)	28	35	38		40
pension	{none, ret-allw, empl-cntr}	none	?	?		?
standby pay	percentage	?	13%	?		?
shift-work supplement	percentage	?	5%	4%		4
education allowance	{yes, no}	yes	?	?		?
statutory holidays	(number of days)	11	15	12		12
vacation	{below-avg, avg, gen}	avg	gen	gen		avg
long-term disability assistance	{yes, no}	no	?	?		yes
dental plan contribution	{none, half, full}	none	?	full		full
bereavement assistance	{yes, no}	no	?	?		yes
health plan contribution	{none, half, full}	none	?	full		half
acceptability of contract	{good, bad}	bad	good	good		good

Types of attributes

- Based on variable type, practical data mining systems divide attributes into two (three) types.
- **Categorical** (enumerated, discrete) corresponding to nominal, binary and ordinal variables (names, symbols)
- **Continuous** (numeric) corresponding to integer, interval-scaled and ratio-scaled variables
- **Ignore** corresponding to variables that are of no significance for the application

Numeric Attributes

- **Numeric attribute:** It has a real-valued or integer-valued domain
- **Interval-scaled attribute:** It may be numeric in form, but the numeric values have no mathematical interpretation (nominal attributes)
- **Ratio-scaled attribute**: It is numeric in form and arithmetic with variables is meaningful

Categorical Attributes

- **Categorical attribute**: It has a set-valued domain composed of a set of symbols.
- **Nominal attribute**: Values in the domain are unordered, and thus only equality comparisons (is one value equal to another?) are meaningful.
- **Ordinal attribute**: Values are ordered, and thus both equality comparisons and inequality comparisons (is one value less than or greater than another?) are allowed.

Nominal Variables

- A variable used to put objects into categories, e.g. the name or color of an object.
- A nominal variable may be numerical in form, but the numerical values have no mathematical interpretation. They are simply labels.
- A **classification** can be viewed as a nominal variable which has been designated as of particular importance.
- A **binary variable** is a special case of a nominal variable that takes only two possible values: true or false, 1 or 0 etc.

Ordinal Variables

- Ordinal variables are similar to nominal variables.
- An ordinal variable has values that can be arranged in a meaningful order, e.g. small, medium, large.

Integer Variables

- Integer variables are ones that take values that are genuine integers.
- For example ‘number of children’. Unlike nominal variables that are numerical in form, arithmetic with integer variables is meaningful (1 child + 2 children = 3 children etc.).

Interval-scaled Variables

- Interval-scaled variables are variables that take numerical values which are measured at equal intervals from a zero point or origin.
- The origin does not imply a true absence of the measured characteristic.
- Celsius temperature scale. To say that one temperature measured in degrees Celsius is greater than another or greater than a constant value such as 25 is clearly meaningful, but to say that one temperature measured in degrees Celsius is twice another is meaningless.

Ratio-scaled Variables

- Ratio-scaled variables are similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic.
- Kelvin temperature. In the former case the zero value corresponds to the lowest possible temperature ‘absolute zero’, so a temperature of 20 degrees Kelvin is twice one of 10 degrees Kelvin.

Computation

	Nominal	**Ordinal**	**Interval**	**Ratio**
frequency distribution	Yes	Yes	Yes	Yes
median and percentiles	No	Yes	Yes	Yes
add or subtract	No	No	Yes	Yes
mean or deviation	No	No	Yes	Yes
ratio or coefficient of variation	No	No	No	Yes

Data Cleaning

- For some applications, the hardest task may be to transform the data into a standard form in which it can be analyzed!!!

- **Noise**: Usage of the term *noise* varies. A *noisy value* to mean one that is valid for the dataset, but is incorrectly recorded (error in the data). Noise is a perpetual problem with real-world data.

Examples

- A numeric variable may only take six different values.
- All the values of a variable may be identical.
- All the values of a variable except one may be identical.
- Some values are outside the normal range of the variable.
- Some values occur an abnormally large number of times.

Missing Data

- In many real-world datasets data values are not recorded for all attributes.
- Some attributes that are not applicable for some instances.
- There are attribute values that should be recorded but are missing:
  - a malfunction of the equipment used to record the data
  - a data collection form to which additional fields were added after some data had been collected
  - information that could not be obtained

Missing Data: Solution

- Discard Instances
  - Delete all instances where there is at least one missing value and use the remainder.
- Replace by Most Frequent/Average Value
  - To estimate each of the missing values using the values that are present in the dataset (most frequent, average).

Reducing the Number of Attributes

- For some datasets there can be substantially more attributes than there are instances, perhaps as many as 10 or even 100 to one…

- When the number of attributes becomes large, there is always a risk that the results obtained will have only superficial accuracy and will actually be less reliable than if only a small proportion of the attributes were used; a case of ‘more means less’.

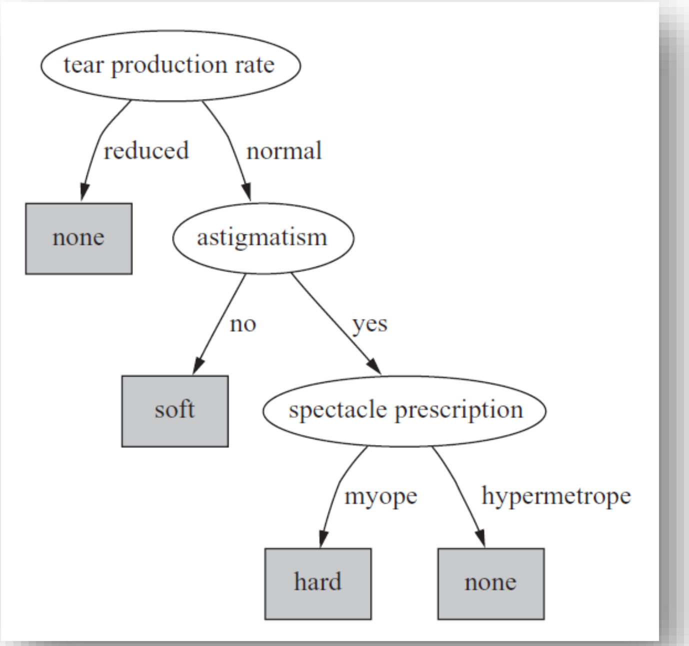
- The term *feature reduction* or *dimension reduction* is generally used for this process.

Contact Lens Data

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Decision Tree (Informally)

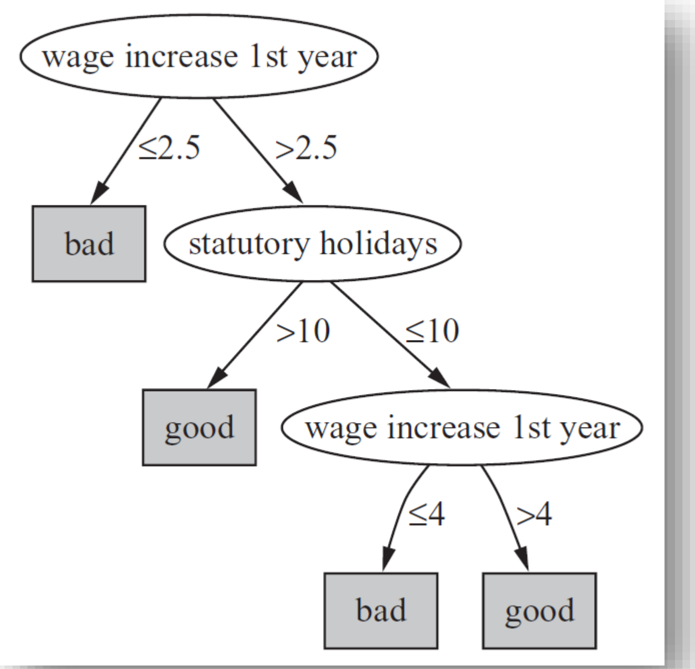
```
If tear production rate = reduced then recommendation = none.
If age = young and astigmatic = no and tear production rate = normal
   then recommendation = soft
If age = pre-presbyopic and astigmatic = no and tear production
   rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope and
   astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no and
   tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = young and astigmatic = yes and tear production rate = normal
   then recommendation = hard
If age = pre-presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
```

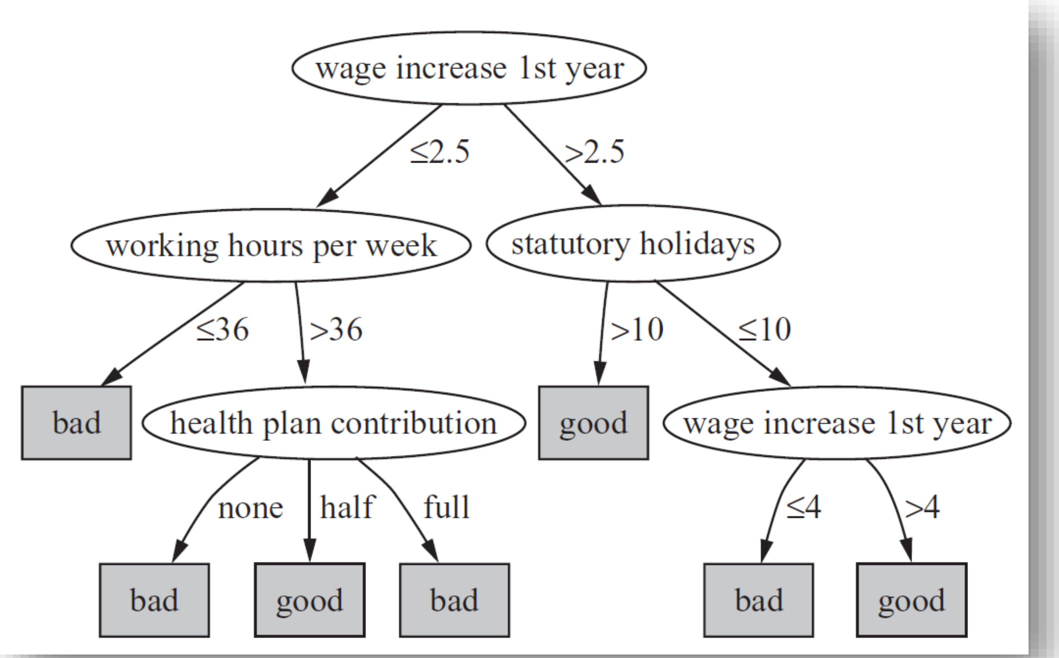


Labor Negotiations Data

Attribute	Type	1	2	3	...	40
duration	(number of years)	1	2	3		2
wage increase 1st year	percentage	2%	4%	4.3%		4.5
wage increase 2nd year	percentage	?	5%	4.4%		4.0
wage increase 3rd year	percentage	?	?	?		?
cost-of-living adjustment	{none, tcf, tc}	none	tcf	?		none
working hours per week	(number of hours)	28	35	38		40
pension	{none, ret-allw, empl-cntr}	none	?	?		?
standby pay	percentage	?	13%	?		?
shift-work supplement	percentage	?	5%	4%		4
education allowance	{yes, no}	yes	?	?		?
statutory holidays	(number of days)	11	15	12		12
vacation	{below-avg, avg, gen}	avg	gen	gen		avg
long-term disability assistance	{yes, no}	no	?	?		yes
dental plan contribution	{none, half, full}	none	?	full		full
bereavement assistance	{yes, no}	no	?	?		yes
health plan contribution	{none, half, full}	none	?	full		half
acceptability of contract	{good, bad}	bad	good	good		good

Decision Trees





Tools and Datasets

- WEKA - http://www.cs.waikato.ac.nz/ml/weka/
- Repository - http://archive.ics.uci.edu/ml/datasets.html